IBM Research 数理科学技術研究への取り組み
～どうする、いつする、だれがする～

Vice President,
Business Analytics and
Mathematical Research
IBM Research

Dr. Robert S. Sutor

ビジネス関連のメディアに目を向けると、
「アナリティクス」という言葉が繰り返し出てきます。
しかしそこでは、数学、統計学、データ集約などの趣旨で
漠然と使われており、多くの場合、明確な定義がなされていません。
そのため、アナリティクスを適用できる問題を識別することも、
問題の解決に必要な人材が確保されているかどうかを
判断することにも困難が伴います。
本稿では、アナリティクスとその関連分野について概観し、
最先端の研究にはどのようなものがあり、
どういった人材と連携すべきかについて理解を深めてもらいたいと思います。
ビッグデータとアナリティクスとを区別することから始め、
次いで IBM Research での私自身のグループの研究分野の
いくつかについて説明します。
最後に、われわれが Smarter Workforce と呼んでいる
人材の最適活用に関する最新の研究内容についても取り上げます。

ビッグデータ

これは今日使用されている最も包括的な用語ですが、それゆえによく分からない言葉となっています。多種多様な大量のデータが瞬時に生まれ、重要で正確な情報とそうでない情報を区別することが難しい状況を想像してみましょう。ここには、Volume（容量）、Variety（種類）、Velocity（スピード）、Veracity（正確さ）という、いわゆる「4つのV」が表すビッグデータの特徴が示されています。

ビッグデータが持つこの4つの側面は、情報を処理するときに考慮すべきポイントを端的に表していると思います。データが、自分で管理している、あるいはアクセスできるデータベースにあるとします。大規模な組織では、情報が部門間に分散していることがよくあり、形式上はすべての情報が自分に属していても、それを使用する権限がなかったり、使用方法を知らなかったりということがあります。こうした異なる情報源を簡単に統合してデータを最適な形で使用するということができない場合もあります。

例えば、皆さんが使っている金融機関は、「自動車保険に入っているロバート・スーター」と「住宅ローンを組んでいるロバート・スーター」が同一人物であることを確認できるでしょうか。もし確認できないとすると、会社が保有しているデータを最大限に生かしきれておらず、最良の顧客サービスを提供できていない可能性があると言えるでしょう。

近年ソーシャル・ネットワークやデバイスがもたらすデータが急増しています。手元のスマートフォンやタブレットが、皆さんがどこで何をしているかという情報を生み出していることは容易に想像できます。車やトラックのような大型の機械でも同様のことが行われています。カメラが作り出すデータは防犯や安全確保のほか、医療や小売の分野にも応用されています。

これらの情報すべてを手にしたとして、何をすべきでしょうか。リアルタイムでデータを眺めるか、保存してあとで処理するか、どのくらいの情報量を保存するか。さらに、ある情報を消去する場合、1～2年後にテクノロジーが向上したときにその削除済みの情報が価値を持たないと言えるかどうか。

膨大な量のデータを抱えているとき、データを迅速に処理し活用するにはどうしたら良いでしょうか。技術者はこの問題を解決するためにさまざまな方法を開発してきました。Hadoopとそれに関連するソフトウェアでは、処理を細分化して複数のマシンに分散させ、それぞれの結果を再びまとめることで目的を達成します。

ストリーム処理では、データが発生した、または受け取った時点でまずデータを眺め、何が重要で何がそうでないかを判断し、重要な情報に対して処理を行います。この処理は、顧客の購入履歴のような既存の静的データやTwitterのコメントのような動的データと結び付けて行われることもあります。

ここまで、現在大量のデータが保管されていることと、新たに膨大な量のデータが生み出されつつあること、こうしたデータを処理する高度な技術が存在することを述べてきましたが、以降では、このようなデータがどう扱われているかについて説明していきます。

一般のメディアでは、ビッグデータですべてが尽きるような言い方、すなわち、ビッグデータという言葉には情報そのものと、その利用シナリオの双方が含まれているように使われています。専門家は多くの場合、ビッグデータという言葉を、私がこれまで述べてきた内容、すなわち情報そのものと基本的な処理技法に限定しています。アナリティクスとはその上に位置する層であり、情報が伝える内容を理解するために使用されています。

アナリティクス

現在、アナリティクスについてWikipediaでは次のように記載されています。

“アナリティクスとは、データに内在する意味のあるパターンを発見し、伝達する技術です。アナリティクスは、豊富に情報が記録されるような分野で特に有用であり、統計学、コンピューター・プログラミング、オペレーションズ・リサーチのような異なる分野の技術を駆使して定量的な分析を実行します。多くの場合、アナリティクスではその知見を伝えるためにデータの可視化が求められます。”

最初の文には、「発見」「伝達」「意味のあるパターン」といった重要な言葉とフレーズが並んでいますが、数ギガバイト、数テラバイト、または数ペタバイトのデータを与えられたとしたら、その意味内容を理解するためにデータをどのように処理したら良いのでしょうか。

例えばこれらのデータが、新しく立ち上げたヘルプデスク業務の顧客満足度に関するアンケート調査結果であると想定します。顧客の満足度が最も高い項目と最も低い項目を自動的に識別できる

でしょうか。また、その項目とお客様が製品を購入した特定の店をひも付けられるでしょうか。また1日のうちでヘルプデスクの顧客満足度が最も高くなる時間帯、最も低くなる時間帯はいつでしょうか。最良のヘルプデスク担当者はどのような特徴を備えているべきでしょうか。これらの情報が意味する内容やデータが結果を示唆する度合い、とるべき行動（必要な場合）について、文書またはビジュアル形式で表現できるでしょうか。

使用するデータを入手したら、不要な部分をフィルターで除去し、残りの部分を整理します。例えば、性別が関係ない場合にはその情報を削除できますが、顧客の姓のスペルの表記は記録全体で統一したほうが良いでしょう。この種の作業は長時間を要することが多く、大規模に自動化されることがあります。しかし一般に、数学、統計学、情報科学の専門知識を必要としません。

適切なデータを確保したら、その数理モデルを作成します。このモデルにより、データの内容を理解し、現在の傾向が続いた場合や何か変革が行われた場合に起こることを予測し、達成目標に応じて結果を最適化することが可能となるでしょう。

先ほどのヘルプデスクを例にとれば、シンプルな最適化であれば、95%の顧客満足度を達成するには特定のスキル・レベルにあるスタッフの20%増員を必要とするなどの提案につながりますし、さらには顧客の役に立つ応答を10分以内に90%以上の確率で行うように主張することもできます。

また、より高度な最適化であれば、製品の販売チャネルの改善方法を検討したり、顧客からのクレームの大きな要因となっている利益の最も少ないチャネルを排除することも可能です。

基本的なモデル化、予測、最適化については、統計、オペレーションズ・リサーチ、データ・マイニング、機械学習、または応用数学の学士号か修士号を持つ人材が1名以上いれば対応可能でしょう（共通モデルに基づく、ごく標準的な作業の場合）。

しかし、新しいアルゴリズムやモデル、技法、確率的／統計的手法を伴うさらに高度な作業では、博士号を持つ人材を必要とする可能性が高くなります。結合された多数のデータ・ソースを複数のモデルと技法で分析する場合にはこのような人材は必須でしょう。通常、アナリティクス・チームでは異なる専門分野の人材を複数擁しています。チーム・メンバーの3分の1が博士号を持ち、残りのメンバーが学士号か修士号を持つというケースは珍しいことではありません。

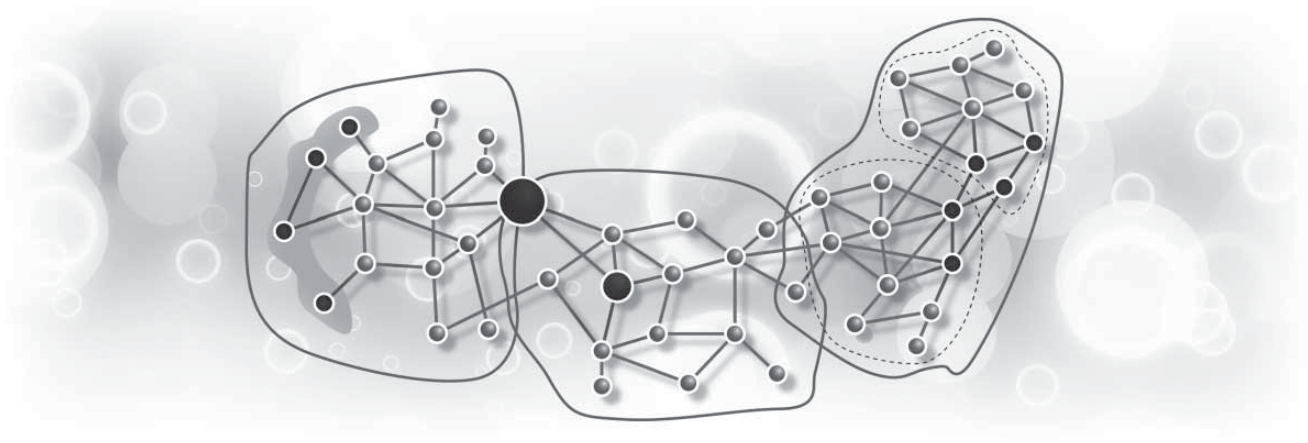
IBM Researchの活動

私が指揮をとっているのは、民間企業としては数学に関して最大規模の研究部門であり、その活動範囲は世界中に広がっています。研究員とソフトウェア・エンジニアは、日本、米国、中国、インド、オーストラリア、ブラジル、イスラエル、アイルランド、スイスにある12の研究所に分散しています。

この部門は「Business Analytics and Mathematical Sciences（ビジネス・アナリティクスおよび数理科学）」という名称ですが、IBM Researchには私たちのほかにもアナリティクスや数学を研究している部門があります。私たちの部門には、核となる数学分野に取り組む科学者が最大規模で集結しており、IBM Researchの同僚やIBMのサービス事業部のスタッフと協力して、さまざまな業界の問題に対して成果を適用しています。また、取り組んでいる各領域では、論文の執筆、学会や業界会議での講演、特許の取得、IBM社内業務の支援、IBM製品の強化、サービス契約を介したお客様に対する直接的な価値提供といった活動も行っています。

IBMのGlobal Technology Outlook（GTO：IBMの技術戦略であり、今後5年から10年先に主流となる技術動向を予測するもの）の2013年版に掲載されたトピックの一つに「ビジュアル・アナリティクス」があります。ビジュアル・アナリティクスは、可視化とは趣を異にします。ビジュアル・アナリティクスを使うことで、データがどのようにふるまい、どう生成されているかについて、対話的なやりかたで、眺めたり、理解したり、触ってみたりできます。ビジュアル・アナリティクスでは、多くの場合、地理データ、運用データ、財務データ、統計データなどのいくつものデータの種類を、ラップトップやタブレット上で簡単に使用できる形式に圧縮します。

ビジュアル・アナリティクスは、対話形式の高度な可視化の仕組みを、洗練されたアナリティクスと結合したものと言えます。以降では、私たちが戦略的イニシアチブとして取り組んでいるアナリティクス分野のうち9つの分野について説明します。私たちの研究対象には、基礎データとモデルの視覚的な表現と統合、バックエンドまたは携帯機器における情報の効率的な保管・処理に必要なクライアント・サーバー・アーキテクチャー、時空間分析におけるユーザー体験の強化（次項で説明）などがあります。



時空間分析

これは空間と時間を同時に分析対象とするアナリティクスを意味する名称です。携帯機器の発展に伴い、昨今とても重要な分野です。携帯機器により、誰がいつどこで何をしているのかを示す情報が手に入るようになり、それが分析の対象となります。応用例としては、伝染病の流行、汚染による影響、天候、ソーシャル・ネットワークの地理的側面が購入に及ぼす影響、営業活動の進捗管理、需要予測などが挙げられます。

分析対象の空間としては二次元と三次元のいずれもありえますが、近年は三次元空間の重要性が高まっています。この重点分野では、より優れたデータ・モデリング方法やそれに基づく正確な予測、得られた洞察を新しいビジュアル・アナリティクス技術を使っていかに表現し伝えるかといった研究も行っています。

イベントの監視、検出、制御

この分野では、急速に発生する多くのイベントについて、正常な動作と異常な動作を識別できることが重要です。例えば、多数の金融トランザクションがひっきりなしに発生する状況においては、発生し続ける膨大なイベントから不正を検出することができます。

同様に、駅構内に複数設置されたビデオカメラから生み出されるデータを解析することで、乗客と駅員の通常の行動や、窃盗や暴力行為、より深刻な犯罪行為が疑われる行動を明らかにすることができます。

相互に作用する複雑なシステムの分析

人間の社会は複雑です。これは都市そのものや、都市の交通網を見れば明らかです。送電網、各種の輸送機関、水の使用、救急医療チームなど、都市の部分的なモデルは作成できるかもしれませんが、そのすべてを正確にモデル化することは極めて困難です。各モデルはそれぞれ複雑化しており、あるモデルの変更は別のモデルの変更をもたらすこともあり、その予測は困難です。このように、各々が相互に作用している複雑なシステムの例は、他にもたくさんあります。

シミュレーションは、こうしたシステムの最適化のためによく使われる技術で、複雑なシステムの各要素を現実に即してシミュレートするには機械学習の手法を活用します。観測データからモデルを構築するための数学的手法は、アナリティクスの予測精度を向上させます。

東京基礎研究所がリードするこの重点分野は、IBM が提唱するSmarter CitiesとSmarter Planetにおける数学的な基盤となっています。

不確実な状況下での意思決定

現実には、ほとんどの出来事が不確実な状況下で発生しています。例えば発電所において、不確定の需要を満たすためにどれだけの電力量が必要かを正確に把握することはできるでしょうか。農作物の生育期の天気は収穫高にどのように影響するのでしょうか。競合他社が類似品を発売した場合、市場での自社製品の売上はどうなるのでしょうか。またそれは、競合品の発売時期によってどのように変わるのでしょうか。

最初から不確実性を考慮に入れておけば、複数の選択肢によって、利益や効率などの指標を最大限に高めることができます。さまざまな方法で不確実性を定量化し分析モデルに取り込むことで、最適化することができるのです。この時使用すべき最適な手法は、不確実性を表すイベントの数と複雑さによって決まります。

収益と価格の最適化

製品やサービスの価格設定と、それによって期待される収益は、需要側の動向に影響される傾向がますます強くなってきています。例えば、ソーシャル・メディアを通じて拡散されたコメントが製品の需要を大きく左右することが挙げられます。そのため、ソーシャル・メディアで影響力を持つ人に対して積極的な低価格設定を行うことで、そのコミュニティー内での自社製品に関する口コミを増やし、販売数量を伸ばすこともできるでしょう。また、過去の購買行動による影響を受けている消費者に対してパーソナライズされた価格を提示することができれば、再び自社から購入する可能性を高めることができるでしょう。

このような需要の創出は、在庫商品と消費者に買ってもらいたい商品を一致させることにつながり、在庫と製造、サプライチェーン全体に影響を与えると言えます。

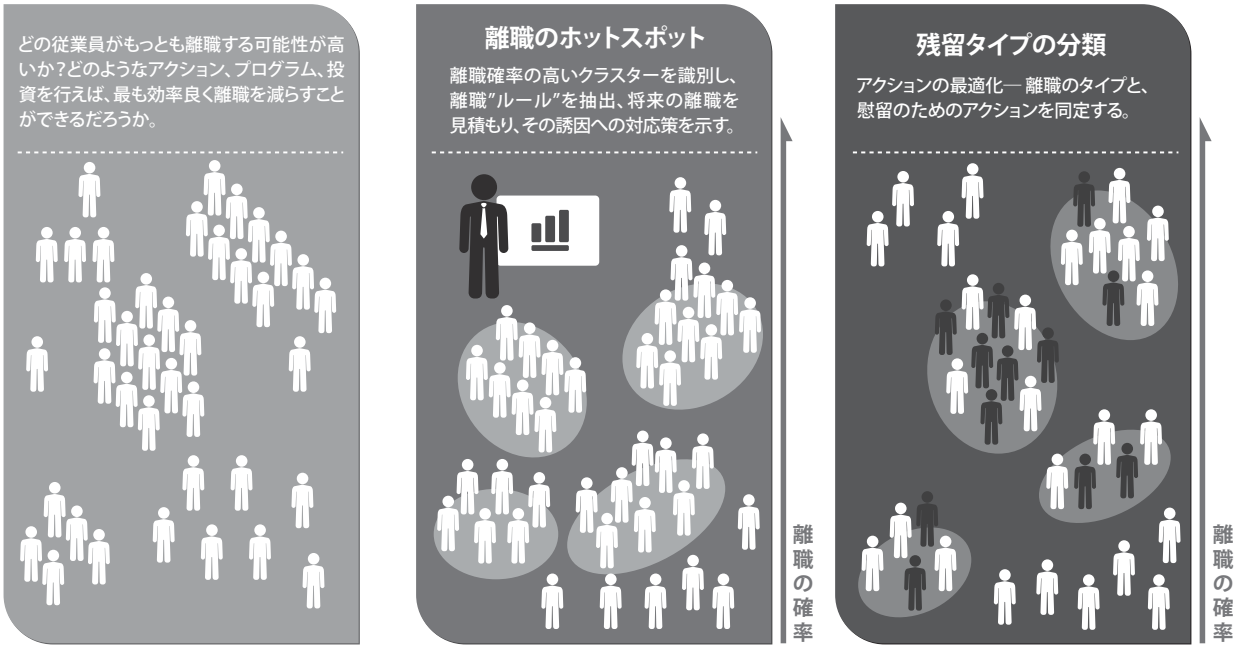
状態基準保全のためのアナリティクス

工場の機械の故障やトラック部品の破損、都市の送水管の漏水はいつ起こるのでしょうか。こうした事態を事前に予測できれば、資産の破損や故障に至る前にメンテナンスを計画し、必要な部品や必要な要員をタイムリーに確保し、事業運営を継続させることができるでしょう。複数の資産が機能しなくなるようなケースにおいては、システム全体が機能し続けるように工程の依存関係なども考慮しながら、どの作業を優先して行うべきかを示すことも可能です。

企業における統合されたオペレーション

この重点分野は、前述の「相互に作用する複雑なシステムの分析」における、組織または企業内のプロセスに特化した応用例とも言えます。例えば鉄鋼会社は、顧客の注文通りに、さまざまな製品を複数の品質等級で製造します。製造に当たっては、在庫資材の最適利用、工作機械の設定とスケジュール、必要最低限のエネルギー使用、必要な品質レベルの維持などが考慮されているでしょう。

各工程を最適化することはできますが、ここでの研究要素は、どうすれば関連作業すべてを最も効果的に行えるかということです。



Smarter Finance

Smarter Finance は、これからの CFO（最高財務責任者）にとって必須の分析ツールになると私は考えています。このソリューションは、運用データと財務データの両方を統合し、リスクやコンプライアンス関連の活動を含めた、組織の全体的な財務方針を最適化します。

Smarter Finance のもう一つの要素が銀行業務への適用です。これには、支店所在地の最適化やクレジット・デフォルトの回収を行う代理店の最適活用などを挙げることができます。

Smarter Workforce

私たちが現在、特に戦略的に重点を置いて取り組んでいるのが、アナリティクスを用いた人材の最適活用、すなわち Smarter Workforce です。私たちは、IBM 社内で 10 年近くにわたってこの取り組みを続けてきました。そして最近、Retention Analytics（リテンション解析）と Survey Analytics（社員満足度解析）の2つをお客様に提供することを発表しました。

Retention Analytics は次のような質問に答えることができます。

- 自社の従業員で最も離職する可能性が高いのは誰か。
- 職務、地域、直近の評価と昇進という観点で、離職する従業員に見られる特徴は何か。
- 辞めた従業員の欠員補充に要するコストはどのくらいか。
- 人員削減による影響を最小化するために、最も残ってほしい従業員への昇給をどう割り当てたら良いのか。

その一方で私たちは、人材の最適活用のための分析技術を、社内施策や財務の分析と関連付けるべく、発展的な研究にも取り組んでいます。例えば、東京の営業スタッフの 10% が 2 週間以内に辞めた場合のこの四半期の収益に及ぼす影響などが挙げられます。

Survey Analytics は、組織内の肯定的ないし否定的な感情を測定します。分析技術は上司の部下に対する知識や認識に取って代わるものではありませんが、Survey Analytics は従業員の意見を入力データとして取り込むことで、ほかの方法では知ることが難しい従業員の情報を見いだすことができます。先にお客様向けヘルプデスクの例を挙げましたが、これは自社の従業員を対象にしたヘルプデスクのようなものです。従業員が最も好む、または最も嫌う特徴や、従業員からの改善に向けた提案をより良く理解するためのものです。

Smarter Workforce は、組織の情報に関する従来型のアナリティクスをソーシャル・データで強化した一例です。現在では、重点分野の多くにソーシャル・メディアのデータ分析が取り込まれています。このアナリティクス自体が豊かな研究分野であり、その正しい実施方法を理解して有益な結果と知見を得ることが求められています。

おわりに

アナリティクスは非常に広範な適用が可能であり、数十年にわたる統計学、応用数学、オペレーションズ・リサーチ、コンピューター・サイエンスの研究の蓄積に基づいています。アナリティクスは、ビッグデータの情報管理における各側面を補完するものです。活用できるデータの量と種類がますます増える中、IBM Research は最先端の研究に引き続き取り組み、データが持つ意味を解明する新しいアルゴリズム、モデル、技法を生み出していきます。そして、この取り組みがお客様と IBM の運用効率と財務実績の向上につながると私たちは信じています。